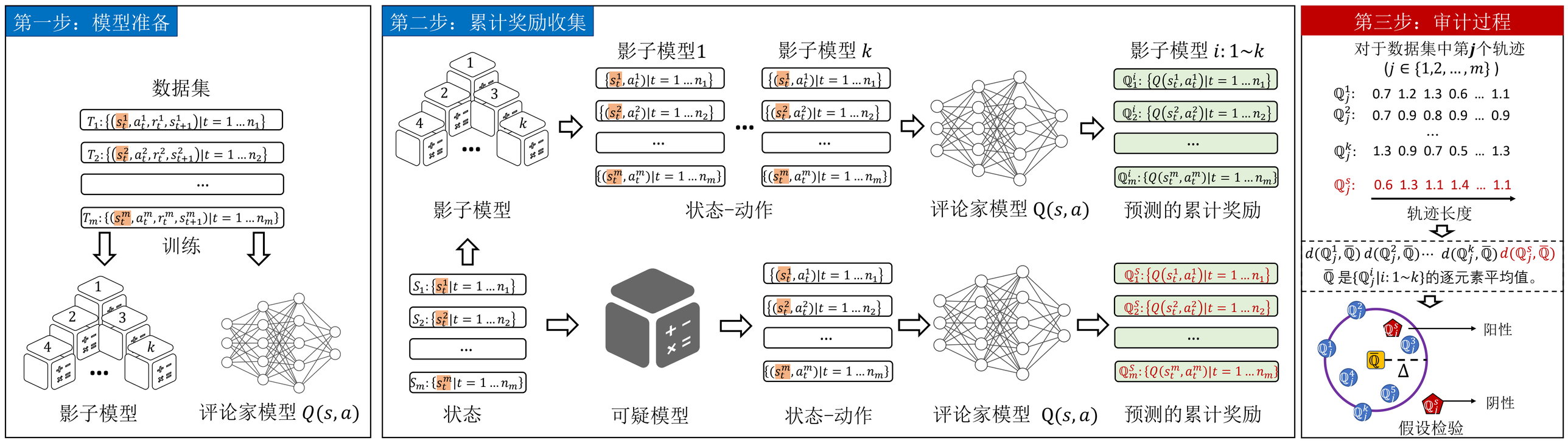
第一步：模型准备
数据集
$T_1: \{(s_t^1, a_t^1, r_t^1, s_{t+1}^1) | t = 1 \dots n_1\}$
$T_2: \{(s_t^2, a_t^2, r_t^2, s_{t+1}^2) | t = 1 \dots n_2\}$
…
$T_m: \{(s_t^m, a_t^m, r_t^m, s_{t+1}^m) | t = 1 \dots n_m\}$
训练
影子模型
评论家模型 $Q(s,a)$
第二步：累计奖励收集
影子模型
影子模型1
$\{(s_t^1, a_t^1) | t = 1 \dots n_1\}$
$\{(s_t^2, a_t^2) | t = 1 \dots n_2\}$
…
$\{(s_t^m, a_t^m) | t = 1 \dots n_m\}$
影子模型 k
$\{(s_t^1, a_t^1) | t = 1 \dots n_1\}$
$\{(s_t^2, a_t^2) | t = 1 \dots n_2\}$
…
$\{(s_t^m, a_t^m) | t = 1 \dots n_m\}$
状态-动作
评论家模型 $Q(s,a)$
影子模型 i: $1 \sim k$
$\mathbb{Q}_1^i: \{Q(s_t^1, a_t^1) | t = 1 \dots n_1\}$
$\mathbb{Q}_2^i: \{Q(s_t^2, a_t^2) | t = 1 \dots n_2\}$
…
$\mathbb{Q}_m^i: \{Q(s_t^m, a_t^m) | t = 1 \dots n_m\}$
预测的累计奖励
$S_1: \{s_t^1 | t = 1 \dots n_1\}$
$S_2: \{s_t^2 | t = 1 \dots n_2\}$
…
$S_m: \{s_t^m | t = 1 \dots n_m\}$
状态
可疑模型
$\{(s_t^1, a_t^1) | t = 1 \dots n_1\}$
$\{(s_t^2, a_t^2) | t = 1 \dots n_2\}$
…
$\{(s_t^m, a_t^m) | t = 1 \dots n_m\}$
状态-动作
评论家模型 $Q(s,a)$
$\mathbb{Q}_1^S: \{Q(s_t^1, a_t^1) | t = 1 \dots n_1\}$
$\mathbb{Q}_2^S: \{Q(s_t^2, a_t^2) | t = 1 \dots n_2\}$
…
$\mathbb{Q}_m^S: \{Q(s_t^m, a_t^m) | t = 1 \dots n_m\}$
预测的累计奖励
第三步：审计过程
对于数据集中第j个轨迹
($j \in \{1,2,\dots,m\}$)
$\mathbb{Q}_j^1$: 0.7 1.2 1.3 0.6 … 1.1
$\mathbb{Q}_j^2$: 0.7 0.9 0.8 0.9 … 0.9
…
$\mathbb{Q}_j^k$: 1.3 0.9 0.7 0.5 … 1.3
$\mathbb{Q}_j^s$: 0.6 1.3 1.1 1.4 … 1.1
轨迹长度
$d(\mathbb{Q}_j^1, \bar{\mathbb{Q}})$ $d(\mathbb{Q}_j^2, \bar{\mathbb{Q}})$ … $d(\mathbb{Q}_j^k, \bar{\mathbb{Q}})$ $d(\mathbb{Q}_j^s, \bar{\mathbb{Q}})$
$\bar{\mathbb{Q}}$ 是$\{\mathbb{Q}_j^i | i: 1 \sim k\}$的逐元素平均值。
阳性
阴性
Δ
假设检验